

**APPLICATION FOR
UNITED STATES PATENT**

in the name of

Nafea Bishara

for

**EFFICIENT IP MULTICAST BRIDGING IN ETHERNET
SWITCHES**

Attorney Docket No. MP0303

EXPRESS MAIL NO.:

Feb. 5, 2004

ER393646163US

EFFICIENT IP MULTICAST BRIDGING IN ETHERNET SWITCHES

Inventor: **Nafea Bishara**

CROSS-REFERENCE TO RELATED APPLICATIONS

- [0001] This application claims the benefit of U.S. Provisional Patent Application Serial No. 60/457,397 entitled "A Method For Efficient IP Multicast Bridging In Ethernet Switches," filed March 24, 2003, the disclosure thereof incorporated by reference herein in its entirety.

BACKGROUND

- [0002] The present invention relates generally to data communications. More particularly, the present invention relates to multicast bridging in network switches.
- [0003] A packet in a packet-switching network such as an Ethernet network can be sent in one of three ways: unicast, multicast, and broadcast. A unicast packet is directed to a single port. A broadcast packet is directed to all of the ports of the network. A multicast packet is directed to a group of the ports of the network. Multicast packets can be either link-layer multicast packets, such as media access control (MAC) multicast packets, or Internet Protocol (IP) multicast packets.
- [0004] Ethernet switches compliant with the IEEE 802.1D-1998 specification are required to switch MAC multicast packets based on the destination MAC address of the packet. Ethernet switches compliant with the IEEE 802.1D-1998, 802.1p and 802.1q specifications are required to switch MAC multicast packets based on the destination MAC address and Virtual Local Area Network (LAN) Identifier (VLAN ID) of the packet. Such switches comprise forwarding databases (FDB) that associate each MAC address (or MAC address and VLAN ID combination) with one or more of the ports of the switch.
- [0005] However, most multicast packets are IP multicast packets, which should be flooded according to the source IP address and IP multicast destination (also referred to as IP multicast group). The Internet Engineering Task Force (IETF) mandates that IP multicast traffic should be encapsulated in a link-layer MAC multicast packet when sent over Ethernet. Unfortunately, Ethernet switches handle IP multicast packets

inefficiently, without regard to the source IP address, according to the MAC destination address only. And because a MAC destination address can map to more than one IP multicast group (32 unique IP multicast addresses map to each MAC multicast address), an IP multicast packet is often flooded to more than one IP multicast group, causing unnecessary network traffic, security breaches, and switch workload.

[0006] For example, consider a switch having four ports p1, p2, p3, and p4. Assume that two IP multicast groups, IPMG1 and IPMG2, are mapped to a single MAC multicast destination address MMDA, that IP multicast group IPMG1 is mapped to ports p1 and p2, and that IP multicast group IPMG2 is mapped to ports p3 and p4. Therefore MAC multicast destination address MMDA is mapped to all four ports. When an IP multicast packet arrives for IP multicast group IPMG1, it is flooded not only to ports p1 and p2, but also to ports p3 and p4. Similarly, IP multicast packets for IP multicast group IPMG2 are flooded to all four ports.

[0007] A similar problem occurs in Ethernet switches that comply with the Internet Group Membership Protocol (IGMPv3). According to the IGMP protocol, a port can ask to receive all of the traffic sent from an IP source address to an IP multicast destination address. However, because Ethernet switches forward IP multicast packets without regard to the source IP address, the port will receive all of the traffic sent to that IP multicast destination address from any IP source address, thus unnecessarily burdening the port and the network.

SUMMARY

[0008] In general, in one aspect, the invention features a method, apparatus, and computer program. The apparatus comprises a plurality of ports each adapted to receive Ethernet packets; and a data-link layer switch controller, when one of the Ethernet packets comprises an Internet protocol (IP) multicast packet comprising an IP multicast destination address and an IP source address, to select one or more of the ports based upon the IP multicast destination address and the IP source address; wherein the selected one or more ports transmit the Ethernet packet.

[0009] Particular implementations can include one or more of the following features. The IP multicast packet comprises a virtual local area network identifier (VLAN ID); and the data-link layer switch controller, is further to select the one or more of the ports based upon the IP multicast destination address, the IP source address, and the

VLAN ID. The apparatus further comprises a memory to store associations between IP addresses and the ports; wherein, to select one or more of the ports based upon the IP multicast destination address and the IP source address, the data-link layer switch controller is further to select the one or more of the ports based upon the associations stored in the memory. To select one or more of the ports based upon the IP multicast destination address and the IP source address, the data-link layer switch controller is further to identify one of the associations stored in the memory based on the IP multicast destination address and the IP source address; and to confirm the association is an association between an IP address and the ports. To identify one of the associations stored in the memory based on the IP multicast destination address and the IP source address, the data-link layer switch controller is further to generate a key based on the IP multicast destination address and the IP source address; and to identify the one of the associations based on the key. To confirm the association is an association between an IP address and the ports, the data-link layer switch controller is further to determine whether the association is marked as an IP multicast association. To determine whether the association is marked as an IP multicast association, the data-link layer switch controller is further to determine whether a flag stored in the memory and corresponding to the association is set. The data-link layer switch controller, when the data-link layer switch controller cannot identify one of the associations stored in the memory based on the IP multicast destination address and the IP source address, is further to generate a message requesting the creation of an association for the IP multicast destination address and the IP source address. The data-link layer switch controller, when the data-link layer switch controller cannot identify one of the associations stored in the memory based on the IP multicast destination address and the IP source address, is further to transmit the Ethernet packet from the ports as destination unknown. The apparatus further comprises a central processing unit to create the association for the IP multicast destination address and the IP source address. The data-link layer switch controller, when one of the Ethernet packets comprises a Media Access Control (MAC) multicast packet comprising a MAC multicast destination address and does not comprise an IP multicast packet, is further to select one or more of the ports based upon the MAC multicast destination address; wherein the selected one or more ports transmit the Ethernet packet. The MAC multicast packet comprises a virtual local area network identifier (VLAN ID); and the data-link layer switch controller, is further to select the one or more of the ports based upon the MAC multicast destination address and the VLAN ID. The apparatus further comprises a memory to store associations between

MAC addresses and the ports; wherein, to select one or more of the ports based upon the MAC multicast destination address, the data-link layer switch controller is further to select the one or more of the ports based upon the associations stored in the memory. To select one or more of the ports based upon the MAC multicast destination address, the data-link layer switch controller is further to identify one of the associations stored in the memory based on the MAC multicast destination address; and to confirm the association is an association between a MAC address and the ports. To identify one of the associations stored in the memory based on the MAC multicast destination address, the data-link layer switch controller is further to generate a key based on the MAC multicast destination address; and to identify the one of the associations based on the key. To confirm the association is an association between a MAC address and the ports, the data-link layer switch controller is further to determine whether the association is marked as a MAC multicast association. To determine whether the association is marked as a MAC multicast association, the data-link layer switch controller is further to determine whether a flag stored in the memory and corresponding to the association is clear. The apparatus further comprises a memory to store a bridge table comprising a plurality of entries each identifying one or more of the ports and addressable by a key; wherein, to select one or more of the ports based upon the IP multicast destination address and the IP source address, the data-link layer switch controller is further to generate the key based upon the IP multicast destination address and the IP source address; wherein, to select one or more of the ports based upon the MAC multicast destination address, the data-link layer switch controller is further to generate the key based upon the MAC multicast destination address; and wherein the selected ports are the ports identified by the bridge table entry addressed by the key. An integrated circuit comprises the apparatus. An Ethernet switch comprises the apparatus.

[0010] The details of one or more implementations are set forth in the accompanying drawings and the description below. Other features will be apparent from the description and drawings, and from the claims.

DESCRIPTION OF DRAWINGS

[0011] FIG. 1 shows an Ethernet switch according to a preferred embodiment of the present invention.

[0012] FIG. 2 shows a forwarding process for the switch of FIG. 1 according to a preferred embodiment of the present invention.

[0013] The leading digit(s) of each reference numeral used in this specification indicates the number of the drawing in which the reference numeral first appears.

DETAILED DESCRIPTION

[0014] Embodiments of the present invention comprise a data-link layer (that is, Open Systems Interconnection (OSI) layer 2) switch controller capable of flooding Ethernet packets encapsulating Internet Protocol (IP) multicast packets based on the IP multicast destination address and IP source address. In contrast to network layer (that is, OSI layer 3) and multi-layer switch controllers, data-link layer switch controllers do not execute network layer protocols, as is well-known in the relevant arts. Further, while network layer and multi-layer switch controllers require separate dedicated network-layer forwarding databases, a link-layer switch controller requires only a bridge table. Embodiments of the data-link layer switch controllers according to the present invention are able to flood Ethernet packets encapsulating IP multicast packets based on the IP multicast destination address and IP source address using the same bridge table that is used for Ethernet bridging.

[0015] When an Ethernet packet is received, the switches of the present invention determine whether the Ethernet packet comprises an IP multicast packet. If so, the switch determines whether the bridge table in the switch contains an entry for the IP multicast destination address and IP source address. If so, the switch floods the Ethernet packet according to that entry. If the IP multicast packet also comprises a virtual local area network identifier (VLAN ID), the switch floods the Ethernet packet according to the IP multicast destination address, the IP source address, and the VLAN ID. But if the bridge table does not contain an entry for the IP multicast destination address and IP source address, the switch optionally sends a message to the central processing unit (CPU) in the switch to request that an entry in the bridge table be created.

[0016] The Ethernet switch is also capable of flooding Ethernet packets encapsulating Media Access Control (MAC) multicast packets that do not encapsulate IP multicast packets based on the MAC multicast destination address in the MAC multicast packet. When an Ethernet packet is received, the switch determines whether the Ethernet packet comprises a MAC multicast packet that does not encapsulate a IP multicast packet. If so, the switch determines whether the bridge table in the switch contains an entry for the MAC multicast destination address in the MAC multicast packet. If so, the switch floods the Ethernet packet according to that entry. If the

MAC multicast packet also comprises a virtual local area network identifier (VLAN ID), the switch floods the Ethernet packet according to the MAC multicast destination address and the VLAN ID.

[0017] FIG. 1 shows an Ethernet switch 100 according to a preferred embodiment of the present invention. Ethernet switch 100 comprises a switch 102, which can be fabricated as a single integrated circuit, and a central processing unit (CPU) 104. Switch 102 comprises a controller 112 and a CPU interface 106 to permit controller 112 to communicate with CPU 104. Switch 102 also comprises a plurality of ports 114A through 114N for exchanging Ethernet packets of data with a network 116 under the control of controller 112 and according to the contents of a bridge table 110 stored in a memory 108.

[0018] FIG. 2 shows a forwarding process 200 for the controller 112 of the switch 102 of FIG. 1 according to a preferred embodiment of the present invention. Process 200 begins when switch 102 receives an Ethernet packet (step 202). Controller 112 determines whether the Ethernet packet comprises an IP multicast packet (step 204) according to conventional techniques, preferably by examining the contents of the packet for a known bit pattern that identifies the packet as an IP multicast packet.

[0019] If the Ethernet packet comprises an IP multicast packet, then controller 112 generates a key based on the IP multicast destination address, IP source address, and if present, VLAN ID in the IP multicast packet and performs a lookup (step 206) on bridge table 110 using the key, which can be generated according to conventional techniques, for example by hashing the IP multicast destination address, IP source address, and if present, VLAN ID using a hash function.

[0020] Controller 112 then confirms that the entry in bridge table 110 indicated by the key is an IP multicast entry (step 208). Each entry in bridge table 110 includes an IP multicast flag that, if set, marks the entry as an IP multicast entry. Each IP multicast entry contains an association between an IP multicast destination address, an IP source address, an optional VLAN ID, and a port indicator that identifies one or more of the ports 114. The port indicator can be a vector comprising a bit representing each port 114, a list of identifiers of one or more ports 114, a pointer to such a port vector or port list, or the like.

[0021] If in step 208 controller 112 finds the IP multicast flag in the entry is set, then controller 112 determines whether the IP multicast destination address, the IP source address, and if present, VLAN ID in the entry match the IP multicast destination

address, IP source address, and if present, VLAN ID in the IP multicast packet (step 210). If they match, then controller 112 floods the Ethernet packet according to the port indicator in the entry (step 212), and process 200 is done (step 214).

[0022] However, if in step 208 the IP multicast flag is clear, or if in step 210 the IP multicast destination address, the IP source address, and if present, VLAN ID in the entry do not match the IP multicast destination address, IP source address, and if present, VLAN ID in the IP multicast packet, controller 112 performs another lookup (step 206) on bridge table 110 based on the lookup algorithm. If the lookup algorithm finishes with no entry matched (step 209), then controller 112 floods the Ethernet packet as destination address unknown (step 216) and optionally generates a message to CPU 104 requesting the creation of an entry in bridge table 110 for the IP multicast destination address, IP source address, and if present, VLAN ID (step 218). Of course, other techniques can be used for notifying CPU 104 and programming bridge table 110. In response, CPU 104 creates such an entry in bridge table 110, and sets the IP multicast flag for the entry. Then process 200 is done (step 214).

[0023] However, if at step 204 the Ethernet packet does not comprise an IP multicast packet, controller 112 determines whether the Ethernet packet comprises a Media Access Control (MAC) multicast packet (step 220) according to conventional techniques, preferably by examining the contents of the packet for a known bit pattern that identifies the packet as a MAC multicast packet.

[0024] If the Ethernet packet comprises a MAC multicast packet, then controller 112 generates a key based on the MAC multicast destination address, and if present, VLAN ID in the MAC multicast packet and performs a lookup (step 222) on bridge table 110 using the key, which can be generated according to conventional techniques, for example by hashing the MAC multicast destination address, and if present, VLAN ID using a hash function.

[0025] Controller 112 then confirms that the entry in bridge table 110 indicated by the key is not an IP multicast entry (step 224) by testing the IP multicast flag. Each MAC multicast entry contains an association between a MAC multicast destination address, an optional VLAN ID, and a port indicator that identifies one or more of the ports 114. The port indicator can be a vector comprising a bit representing each port 114, a list of identifiers of one or more ports 114, a pointer to such a port vector or port list, or the like.

[0026] If in step 224 controller 112 finds the IP multicast flag in the entry is clear, then controller 112 determines whether the MAC multicast destination address, and if present, VLAN ID in the entry match the MAC multicast destination address, and if present, VLAN ID in the MAC multicast packet (step 226). If they match, then controller 112 floods the Ethernet packet according to the port indicator in the entry (step 228), and process 200 is done (step 214).

[0027] However, if in step 224 the IP multicast flag is set, or if in step 226 the MAC multicast destination address, and if present, VLAN ID in the entry do not match the MAC multicast destination address, and if present, VLAN ID in the MAC multicast packet, controller 112 performs another lookup (step 222) on bridge table 110 based on the lookup algorithm. If the lookup algorithm finishes with no entry matched (step 225), then controller 112 floods the Ethernet packet as destination address unknown (step 230) and generates a new MAC multicast entry in bridge table 110 according to conventional methods, ensuring that the IP multicast flag for the entry is clear (step 232). Then process 200 is done (step 214).

[0028] If at step 220 the Ethernet packet comprises neither an IP multicast packet nor a MAC multicast packet, controller 112 floods the Ethernet packet normally, according to conventional techniques (step 234).

[0029] The invention can be implemented in digital electronic circuitry, or in computer hardware, firmware, software, or in combinations of them. Apparatus of the invention can be implemented in a computer program product tangibly embodied in a machine-readable storage device for execution by a programmable processor; and method steps of the invention can be performed by a programmable processor executing a program of instructions to perform functions of the invention by operating on input data and generating output. The invention can be implemented advantageously in one or more computer programs that are executable on a programmable system including at least one programmable processor coupled to receive data and instructions from, and to transmit data and instructions to, a data storage system, at least one input device, and at least one output device. Each computer program can be implemented in a high-level procedural or object-oriented programming language, or in assembly or machine language if desired; and in any case, the language can be a compiled or interpreted language. Suitable processors include, by way of example, both general and special purpose microprocessors. Generally, a processor will receive instructions and data from a read-only memory and/or a random access memory. Generally, a computer will include one or more

mass storage devices for storing data files; such devices include magnetic disks, such as internal hard disks and removable disks; magneto-optical disks; and optical disks. Storage devices suitable for tangibly embodying computer program instructions and data include all forms of non-volatile memory, including by way of example semiconductor memory devices, such as EPROM, EEPROM, and flash memory devices; magnetic disks such as internal hard disks and removable disks; magneto-optical disks; and CD-ROM disks. Any of the foregoing can be supplemented by, or incorporated in, ASICs (application-specific integrated circuits).

[0030] A number of implementations of the invention have been described. Nevertheless, it will be understood that various modifications may be made without departing from the spirit and scope of the invention. Accordingly, other implementations are within the scope of the following claims.